Optimization toolbox for deep learning:

The convergence analysis of Gradient Descent & Gradient Flow

Yushun Zhang

The Chinese University of Hong Kong, shenzhen, China

March 25, 2021

Mainly based on:
Matus Telgarsky. Deep learning theory lecture notes. 2021-01-18 v0.0-e018a871 (pre-alpha)
Chapter 9, 10, 11
Andre Milzarek. Optimization Theory and Algorithms CIE / DDA 6010 – Fall 2020/21

Introduction

Goal:

$$\min_{w} \widehat{\mathcal{R}}(w) \tag{1}$$

where $\widehat{\mathcal{R}}(w) := n^{-1} \sum_i \ell\left(y_i f\left(x_i; w\right)\right)$

- We will cover primarily first-order methods, e.g, GD.
- we'll cover classical inequalities when $\widehat{\mathcal{R}}(w)$ is:
  - a) smooth and nonconvex,
  - b) smooth and convex,
  - c)strongly convex.

Introduction

Most of our contents will be based on first order methods:

- We will cover primarily first-order methods, namely gradient descent:

$$w_{t+1} := w_t - \eta_t \nabla \widehat{\mathcal{R}}\,(w_t)\,(\eta_t \text{ sufficiently small})$$

as well as the gradient flow

$$\frac{\mathrm{d}w}{\mathrm{d}t} = \dot{w}(t) = -\nabla \widehat{\mathcal{R}}(w(t))$$

Warm-up question: How are these two related?

Introduction

- velocity of w: $\frac{\mathrm{d}w}{\mathrm{d}t} = \dot{w}(t) = -\nabla\widehat{\mathcal{R}}(w(t))$
- if at time t we are at point w_t and know our velocity is $\dot{w}(t)$, where should we go next? The velocity vector tells us where we will approximately be in the near future:
- $w_{t+\delta} \approx w_t + \delta\dot{w}_t = w_t + \delta - \nabla\widehat{\mathcal{R}}(w(t))$
- Define a discretized sequence: $w_k := w_{\delta k}$
- then we obtain an algorithm in discrete time: $w_{k+1} = w_k + \delta F\left(w_k\right)$

Thus, we can view algorithms in discrete time as a discretization of dynamics in continuous time, or dynamics as the continuous-time limit $(\delta \to 0)$ of algorithms.

Outline

Smooth and nonconvex case

Definition 1.
We say " $\widehat{\mathcal{R}}$ is β -smooth" to mean β -Lipschitz gradients:

$$\|\nabla\widehat{\mathcal{R}}(w) - \nabla\widehat{\mathcal{R}}(v)\| \leq \beta\|w - v\|$$

Lemma 1.
Descent lemma: *When $\widehat{\mathcal{R}}$ is β -smooth, we have:*

$$\widehat{\mathcal{R}}(v) \leq \widehat{\mathcal{R}}(w) + \langle\nabla\widehat{\mathcal{R}}(w), v - w\rangle + \frac{\beta}{2}\|v - w\|^2$$

Smooth and nonconvex case

Proof of Descent lemma:

By the Fundamental theorem of calculus:

$$
\begin{aligned}
&|\widehat{\mathcal{R}}(v) - \widehat{\mathcal{R}}(w) - \langle \nabla \widehat{\mathcal{R}}(w), v - w \rangle| \\
&= |\int_0^1 \langle \nabla \widehat{\mathcal{R}}(w + t(v - w)), v - w \rangle \mathrm{d}t - \langle \nabla \widehat{\mathcal{R}}(w), v - w \rangle| \\
&\leq \int_0^1 |\langle \nabla \widehat{\mathcal{R}}(w + t(v - w)) - \nabla \widehat{\mathcal{R}}(w), v - w \rangle| \mathrm{d}t \\
&\leq \int_0^1 \|\nabla \widehat{\mathcal{R}}(w + t(v - w)) - \nabla \widehat{\mathcal{R}}(w)\| \cdot \|v - w\| \mathrm{d}t \\
&\leq \int_0^1 t\beta \|t(v - w) + w - w\| \|v - w\| \ \mathrm{d}t \\
&\leq \int_0^1 t\beta \|v - w\|^2 \ \mathrm{d}t \\
&= \frac{\beta}{2} \|v - w\|^2
\end{aligned}
$$

$\square$

Smooth and nonconvex case

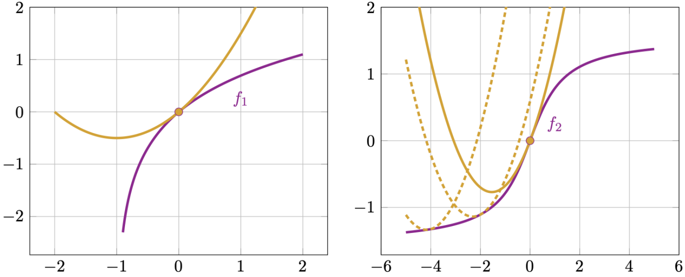

Figure 5.4: Plot of $f_1(x) = \log(1+x)$ and $f_2(x) = \arctan(x)$ and of the quadratic upper surrogate functions $q_1(y) = y + \frac{1}{2}y^2$ and $q_2(y) = y + \frac{3\sqrt{3}}{16}y^2$ at $x = 0$. (The Lipschitz constant of the derivatives $f_1'(x) = (1+x)^{-1}$ and $f_2'(x) = (1+x^2)^{-1}$ is given by $L = 1$ and $L = 3\sqrt{3}/8$, respectively).

Smooth and nonconvex case

Remark 1.
Consider gradient iteration $w' = w - \frac{1}{\beta}\nabla\widehat{\mathcal{R}}(w)$, *then the descent lemma implies:*

$$\widehat{\mathcal{R}}\left(w'\right) \leq \widehat{\mathcal{R}}(w) - \langle\widehat{\mathcal{R}}(w), \widehat{\mathcal{R}}(w)/\beta\rangle + \frac{1}{2\beta}\|\widehat{\mathcal{R}}(w)\|^2 = \widehat{\mathcal{R}}(w) - \frac{1}{2\beta}\|\nabla\widehat{\mathcal{R}}(w)\|^2 \tag{2}$$

- *we can guarantee gradient descent does not increase the objective.*
- *This inequality will occur a lot.*

Smooth and nonconvex case

Remark 2.
Consider gradient iteration $w' = w - \eta\nabla\widehat{\mathcal{R}}(w)$, then the descent lemma implies:

$$\begin{aligned}
\widehat{\mathcal{R}}\left(w'\right) &\leq \widehat{\mathcal{R}}(w) + \left\langle \nabla\widehat{\mathcal{R}}(w), w' - w \right\rangle + \frac{\beta}{2}\left\|w' - w\right\|^2 && (3)\\
&= \widehat{\mathcal{R}}(w) - \eta\|\nabla\widehat{\mathcal{R}}(w)\|^2 + \frac{\beta\eta^2}{2}\|\nabla\widehat{\mathcal{R}}(w)\|^2 && (4)\\
&= \widehat{\mathcal{R}}(w) - \eta\left(1 - \frac{\beta\eta}{2}\right)\|\nabla\widehat{\mathcal{R}}(w)\|^2 && (5)
\end{aligned}$$

If we choose η appropriately $(\eta \leq 2/\beta)$ then:

- either we are near a critical point $(\nabla\widehat{\mathcal{R}}(w) \approx 0)$,
- or we can decrease $\widehat{\mathcal{R}}(w)$.

Smooth and nonconvex case

Theorem 1.

Let $(w_i)_{i\geq 0}$ be given by gradient descent on β -smooth $\widehat{\mathcal{R}}(w)$. For stepsize $\eta \leq \frac{2}{\beta}$:

$$\min_{i<t} \|\nabla \widehat{\mathcal{R}}(w)\|^2 \leq \frac{1}{t}\sum_{i<t} \|\nabla \widehat{\mathcal{R}}(w)\|^2 \tag{6}$$

$$\overset{(\textit{Remark 2})}{\leq} \frac{2}{t\eta(2-\eta\beta)} \left(\widehat{\mathcal{R}}\left(w_0\right) - \widehat{\mathcal{R}}\left(w_t\right)\right) \tag{7}$$

$$\leq \frac{2}{t\eta(2-\eta\beta)} \left(\widehat{\mathcal{R}}\left(w_0\right) - \inf_w \widehat{\mathcal{R}}(w)\right) \tag{8}$$

E.g. when $\eta = \frac{1}{\beta}$ we have: $\min_{i<t} \|\nabla \widehat{\mathcal{R}}(w)\|^2 \leq \frac{2\beta}{t} \left(\widehat{\mathcal{R}}\left(w_0\right) - \widehat{\mathcal{R}}\left(w_t\right)\right)$.

Smooth and nonconvex case

Remark 3.
We have no guarantee about the last iterate $\left\|\nabla\widehat{\mathcal{R}}\left(w_t\right)\right\|$ *: we may get near a flat region at some* $i < t$*, but thereafter bounce out. With a more involved proof, we can guarantee we bounce out (J. D. Lee et al. 2016), but there are cases where the time is exponential in dimension.*

Remark 4.
This derivation is at the core of many papers with a "local optimization" (stationary point or local optimum) guarantee for gradient descent.

Smooth and nonconvex case

Remark 5.
The gradient iterate with step size $\frac{1}{\beta}$ is the result of minimizing the quadratic provided by smoothness:

$$
\begin{aligned}
w - \frac{1}{\beta}\nabla\widehat{\mathcal{R}}(w) &= \arg\min_{w'}\left(\widehat{\mathcal{R}}(w) + \left\langle \nabla\widehat{\mathcal{R}}(w), w' - w\right\rangle + \frac{\beta}{2}\left\|w' - w\right\|^2\right) && (9)\\
&= \arg\min_{w'}\left(\left\langle \nabla\widehat{\mathcal{R}}(w), w'\right\rangle + \frac{\beta}{2}\left\|w' - w\right\|^2\right) && (10)
\end{aligned}
$$

This relates to proximal descent and mirror descent generalizations of gradient descent.

Smooth and nonconvex case

Gradient flow (GF) version:

Recall GF: $\dot{w}(t) = -\nabla\widehat{\mathcal{R}}(w(t))$. Using FTC, chain rule, and definition:

$$
\begin{aligned}
\widehat{\mathcal{R}}(w(t)) - \widehat{\mathcal{R}}(w(0)) &\overset{\text{(FTC)}}{=} \int_0^t \dot{\overbrace{\widehat{\mathcal{R}}(w(t))}}\mathrm{d}t && (11)\\
&\overset{\text{(Chain rule)}}{=} \int_0^t \langle \nabla\widehat{\mathcal{R}}(w(s)), \dot{w}(s)\rangle \mathrm{d}s && (12)\\
&\overset{\text{(GF)}}{=} -\int_0^t \|\nabla\widehat{\mathcal{R}}(w(s))\|\mathrm{d}s && (13)\\
&\leq -t \inf_{s\in[0,t]} \|\nabla\widehat{\mathcal{R}}(w(s))\|^2 && (14)
\end{aligned}
$$

Smooth and nonconvex case

Theorem 2.

For the gradient flow:

$$\inf_{s\in[0,t]} \|\nabla\widehat{\mathcal{R}}(w(s))\|^2 \leq \frac{1}{t}(\widehat{\mathcal{R}}(w(0)) - \widehat{\mathcal{R}}(w(t))) \tag{15}$$

Remark 6.

Compare with GD: $\min_{i<t} \|\nabla\widehat{\mathcal{R}}(w)\|^2 \leq \frac{2\beta}{t}\left(\widehat{\mathcal{R}}\left(w_0\right) - \widehat{\mathcal{R}}\left(w_t\right)\right)$

- β *is from step size.*
- *"2" is from the smoothness term in descent lemma. (avoided in GF).*

Smooth and nonconvex case

Discussion:

- All the previous results are based on the assumption: $\nabla \widehat{\mathcal{R}}(w)$ is Lipschitz continuous.
- This may not be true for NNs.
- Yet, people still assume the iterates are bounded so that $\nabla \widehat{\mathcal{R}}(w)$ is Lipschitz continuous.
- How to ensure the boundedness of iterates?
  - For convex problem: Adaptive Gradient Descent without Descent.
  - For nonconvex NNs: Add a regularization term to the loss, aka, weight decay.

Outline

Smooth and convex case

For convex functions, we have subgradient inequalities:

Lemma 3.
Subgradient inequality: *For any* w' *and* w*:* $\widehat{\mathcal{R}}\left(w'\right) \geq \widehat{\mathcal{R}}(w) + \left\langle \nabla \widehat{\mathcal{R}}(w), w' - w \right\rangle$

Theorem 4.
Suppose $\widehat{\mathcal{R}}$ *is* β *-smooth and convex, and* $(w_i) \geq 0$ *given by GD with* $\eta_i := 1/\beta$*, then for any* z*:*

$$\widehat{\mathcal{R}}\left(w_t\right) - \widehat{\mathcal{R}}(z) \leq \frac{\beta}{2t}\left(\left\|w_0 - z\right\|^2 - \left\|w_t - z\right\|^2\right) \tag{16}$$

*: The reference point z allows us to use this bound effectively when $\widehat{\mathcal{R}}$ lacks an optimum, or simply when the optimum is very large. (linear separable logistic regression.)

Smooth and convex case

Proof of Theorem 4:

$$\begin{aligned}
\|w' - z\|^2 &= \|w' - w + w - z\|^2 && (17)\\
&\overset{(\mathrm{GD})}{=} \| - \frac{1}{\beta}\nabla\widehat{\mathcal{R}}(w) + w - z\|^2 && (18)\\
&= \|w - z\|^2 - \frac{2}{\beta}\langle \nabla\widehat{\mathcal{R}}(w), w - z\rangle + \frac{1}{\beta^2}\|\nabla\widehat{\mathcal{R}}(w)\|^2 && (19)\\
&\overset{(1)(2)}{\leq} \|w - z\|^2 + \frac{2}{\beta}(\widehat{\mathcal{R}}(z) - \widehat{\mathcal{R}}(w)) + \frac{2}{\beta}\left(\widehat{\mathcal{R}}(w) - \widehat{\mathcal{R}}(w')\right) && (20)\\
&= \|w - z\|^2 + \frac{2}{\beta}\left(\widehat{\mathcal{R}}(z) - \widehat{\mathcal{R}}(w')\right) && (21)
\end{aligned}$$

where (1): subgradient inequality, (2): Descent lemma.

$\widehat{\mathcal{R}}(w') \leq \widehat{\mathcal{R}}(w) + \langle \nabla\widehat{\mathcal{R}}(w), w' - w\rangle + \frac{\beta}{2}\|w' - w\|^2 = \widehat{\mathcal{R}}(w) - \frac{1}{\beta}\|\nabla\widehat{\mathcal{R}}(w)\|^2 + \frac{1}{2\beta}\|\nabla\widehat{\mathcal{R}}(w)\|^2$ $\square$

Smooth and convex case

Proof continued:

Rearranging and applying $\sum_{i<t}$

$$\frac{2}{\beta}\sum_{i<t}\left(\widehat{\mathcal{R}}\left(w_{i+1}\right)-\widehat{\mathcal{R}}(z)\right)\leq\sum_{i<t}\left(\left\|w_i-z\right\|^2-\left\|w_{i+1}-z\right\|^2\right)$$

The final bound follows by noting $\widehat{\mathcal{R}}\left(w_i\right)\geq\widehat{\mathcal{R}}\left(w_t\right)$, and since the right hand side telescopes.

$$\widehat{\mathcal{R}}\left(w_t\right)-\widehat{\mathcal{R}}(z)\leq\frac{\beta}{2t}\left(\left\|w_0-z\right\|^2-\left\|w_t-z\right\|^2\right)$$

$\square$

Smooth and convex case

We have similar results for gradient flow (GF):

Theorem 5.
Suppose $\widehat{\mathcal{R}}$ is β -smooth and convex, for any z, GF satisfies:

$$\widehat{\mathcal{R}}\left(w(t)\right) - \widehat{\mathcal{R}}(z) \leq \frac{1}{2t}\left(\|w(0) - z\|^2 - \|w(t) - z\|^2\right) \tag{22}$$

*: difference with GD: no β. Are these two results consistent with each other? Yes

Smooth and convex case

Are these two results consistent with each other? Yes

- Suppose $\|\nabla\widehat{\mathcal{R}}(w)\| \approx 1$ for sake of illustration.
- The "distance traveled" by GD: $\|w_t - w_0\| = \left\|\frac{1}{\beta}\sum_i \nabla\widehat{\mathcal{R}}\left(w_i\right)\right\| \leq \sum_i \frac{1}{\beta}\left\|\nabla\widehat{\mathcal{R}}\left(w_i\right)\right\| \approx \frac{t}{\beta}$
- The "distance traveled" by GF is (via Jensen):

$$\begin{aligned}
\|w(t) - w(0)\| &= \left\|\int_0^t \nabla\widehat{\mathcal{R}}(w(s))\mathrm{d}s\right\| = \left\|\frac{1}{t}\int_0^t t\nabla\widehat{\mathcal{R}}(w(s))\mathrm{d}s\right\| \\
&\leq \frac{1}{t}\int_0^t \|t\nabla\widehat{\mathcal{R}}(w(s))\|\mathrm{d}s \approx t
\end{aligned}$$

- So for GD and GF: $\widehat{\mathcal{R}}\left(w(t)\right) - \widehat{\mathcal{R}}(z)$ are of the same order.

Smooth and convex case

Proof of Theorem 5: By the Fundamental Theorem of Calculus (FTC):

$$\frac{1}{2}\|w(t)-z\|_2^2-\frac{1}{2}\|w(0)-z\|_2^2 \overset{\text{(FTC)}}{=} \frac{1}{2}\int_0^t \frac{\mathrm{d}}{\mathrm{d}s}\|w(s)-z\|_2^2 \,\mathrm{d}s \tag{23}$$

$$\overset{\text{(Chain rule)}}{=} \int_0^t \left\langle \frac{\mathrm{d}w}{\mathrm{d}s}, w(s)-z \right\rangle \mathrm{d}s \tag{24}$$

$$\overset{\text{(subgradient inequality)}}{\leq} \int_0^t (\widehat{\mathcal{R}}(z)-\widehat{\mathcal{R}}(w(s)))\mathrm{d}s; \tag{25}$$

Then we have:

$$t\widehat{\mathcal{R}}(w(t))+\frac{1}{2}\|w(t)-z\|_2^2 \overset{(1)}{\leq} \int_0^t \widehat{\mathcal{R}}(w(s))\mathrm{d}s+\frac{1}{2}\|w(t)-z\|_2^2 \tag{26}$$

$$\leq t\widehat{\mathcal{R}}(z)+\frac{1}{2}\|w(0)-z\|_2^2 \tag{27}$$

* (1): $\widehat{\mathcal{R}}(w(t))$ is nonincreasing in t

Smooth and convex case

Some rules of thumb for convex opt (not comprehensive, and there are other ways).

- $\frac{1}{\sqrt{t}}$ uses Lipschitz of $\widehat{\mathcal{R}}$, (thus $\|\nabla\widehat{\mathcal{R}}\| = \mathcal{O}(1)\Big)$ in place of smoothness upper bound on $\|\nabla\widehat{\mathcal{R}}\|$.
- $\frac{1}{t}$ is often from Lipschitz Gradient.
- $\frac{1}{t^2}$ uses "acceleration," which is a fancy momentum inside the gradient.
- $\exp(-\mathcal{O}(t))$ uses strong convexity (or other fine structure on $\widehat{\mathcal{R}}$).
- Stochasticity changes some rates and what is possible, but there are multiple settings and inconsistent terminology.

Outline

Smooth and strongly convex case

Definition 6.

We say that smooth function $\widehat{\mathcal{R}}$ is λ -strongly-convex $(\lambda - \mathbf{sc})$ when

$$\widehat{\mathcal{R}}\left(w'\right) \geq \widehat{\mathcal{R}}(w) + \left\langle \nabla \widehat{\mathcal{R}}(w), w' - w \right\rangle + \frac{\lambda}{2} \left\| w' - w \right\|^2 \tag{28}$$

* Strongly convex function could be nonsmooth, with an alternative definition:
$\widehat{\mathcal{R}}$ is λ -sc iff $\widehat{\mathcal{R}} - \|\cdot\|_2^2/2$ is convex.

Smooth and strongly convex case

Lemma 7.

PL condition: *Suppose $\widehat{\mathcal{R}}$ is λ -sc. Then we have*

$$\forall w \cdot \quad \widehat{\mathcal{R}}(w) - \inf_{v} \widehat{\mathcal{R}}(v) \leq \frac{1}{2\lambda} \|\nabla \widehat{\mathcal{R}}(w)\|^2 \tag{29}$$

Remark 7.

- *Every stationary point is a global min.*
- *Recall descent lemma:* $\frac{1}{2\beta} \left\|\nabla \widehat{\mathcal{R}}\left(w_i\right)\right\|^2 \leq \widehat{\mathcal{R}}\left(w_i\right) - \widehat{\mathcal{R}}\left(w_{i+1}\right)$, *which means every limit point of GD will be a global min.*

Smooth and strongly convex case

Proof of PL condition:

Let w be given, and define the convex quadratic: (By $\lambda - \mathrm{sc}$, $\widehat{\mathcal{R}}(v) \geq Q_w(v)$.)

$$Q_w(v) := \widehat{\mathcal{R}}(w) + \langle \nabla \widehat{\mathcal{R}}(w), v - w \rangle + \frac{\lambda}{2}\|v - w\|^2 \tag{30}$$

which attains its minimum at $\bar{v} := w - \nabla \widehat{\mathcal{R}}(w)/\lambda$. By definition $\lambda - \mathrm{sc}$

$$\inf_v \widehat{\mathcal{R}}(v) \geq \inf_v Q_w(v) = Q_w(\bar{v}) = \widehat{\mathcal{R}}(w) - \frac{1}{2\lambda}\|\nabla \widehat{\mathcal{R}}(w)\|^2 \tag{31}$$

$\square$

Smooth and strongly convex case

Lemma 8.

(weight decay): *Given* $\widehat{\mathcal{R}}_\lambda(w) = \widehat{\mathcal{R}}(w) + \lambda\|w\|^2/2$ *with* $\widehat{\mathcal{R}} \geq 0$, *optimal point* $\bar{w}$ *satisfies*

$$\frac{\lambda}{2}\|\bar{w}\|_2^2 \overset{(1)}{\leq} \widehat{\mathcal{R}}_\lambda(\bar{w}) \overset{(2)}{\leq} \widehat{\mathcal{R}}_\lambda(0) = \widehat{\mathcal{R}}(0) \tag{32}$$

* (1): $\widehat{\mathcal{R}} \geq 0$,, (2): plug in $w = 0$. No convexity used here.

Remark 8.

- *thus it suffices to search over bounded set* $\left\{w \in \mathbb{R}^p : \|w\|^2 \leq 2\widehat{\mathcal{R}}(0)/\lambda\right\}$. *This can often be plugged directly into generalization bounds.*
- *In deep learning, this style of regularization ("weight decay") is indeed used, but it isn't necessary for generalization. (Chiyuan Zhang, rethinking generalization)*

Smooth and strongly convex case

Theorem 9.

Suppose $\widehat{\mathcal{R}}(w)$ *is* $\lambda - \mathrm{sc}$ *and* β *-smooth, and GD is run with step size* $1/\beta$. *Then a* $\mathrm{minimum}\,\bar{w}$ *exists, and*

$$\widehat{\mathcal{R}}\left(w_t\right) - \widehat{\mathcal{R}}(\bar{w}) \leq \left(\widehat{\mathcal{R}}\left(w_0\right) - \widehat{\mathcal{R}}(\bar{w})\right) \exp(-t\lambda/\beta) \tag{33}$$

$$\left\|w_t - \bar{w}\right\|^2 \leq \left\|w_0 - \bar{w}\right\|^2 \exp(-t\lambda/\beta) \tag{34}$$

Remark 9.

β/λ *is often called the condition number, we call the problem is well-conditioned when* $\beta/\lambda \approx 1$.

Smooth and strongly convex case

Proof of Theorem 9:

$$\begin{aligned}
\widehat{\mathcal{R}}\left(w_{i+1}\right) - \widehat{\mathcal{R}}(\bar{w}) &\overset{\text{(Descent lemma)}}{\leq} \widehat{\mathcal{R}}\left(w_i\right) - \widehat{\mathcal{R}}(\bar{w}) - \frac{\left\|\nabla \widehat{\mathcal{R}}\left(w_i\right)\right\|^2}{2\beta} && (35)\\
&\overset{\text{(PL condition)}}{\leq} \widehat{\mathcal{R}}\left(w_i\right) - \widehat{\mathcal{R}}(\bar{w}) - \frac{2\lambda\left(\widehat{\mathcal{R}}\left(w_i\right) - \widehat{\mathcal{R}}(\bar{w})\right)}{2\beta} && (36)\\
&\leq \left(\widehat{\mathcal{R}}\left(w_i\right) - \widehat{\mathcal{R}}(\bar{w})\right)(1-\lambda/\beta) && (37)
\end{aligned}$$

Repeat it for i times: since $\prod_{i<t}(1-\lambda/\beta) \leq \prod_{i<t}\exp(-\lambda/\beta) = \exp(-t\lambda/\beta)$ which gives the first bound. □

Smooth and strongly convex case

Proof of Theorem 9:

$$
\begin{aligned}
\|w' - \bar{w}\|^2 &\overset{\text{(GD)}}{=} \|w - \frac{1}{\beta}\nabla\widehat{\mathcal{R}}(w) - \bar{w}\|^2 && (38)\\
&= \|w - \bar{w}\|^2 + \frac{2}{\beta}\langle\nabla\widehat{\mathcal{R}}(w), \bar{w} - w\rangle + \frac{1}{\beta^2}\|\nabla\widehat{\mathcal{R}}(w)\|^2 && (39)\\
&\overset{((1)\ \&\ (2))}{\leq} \|w - \bar{w}\|^2 + \frac{2}{\beta}\left(\widehat{\mathcal{R}}(\bar{w}) - \widehat{\mathcal{R}}(w) - \frac{\lambda}{2}\|\bar{w} - w\|_2^2\right) && (40)\\
&\quad + \frac{1}{\beta^2}\left(2\beta\left(\widehat{\mathcal{R}}(w) - \widehat{\mathcal{R}}(w')\right)\right) && (41)\\
&= (1 - \lambda/\beta)\|w - \bar{w}\|^2 + \frac{2}{\beta}\left(\widehat{\mathcal{R}}(\bar{w}) - \widehat{\mathcal{R}}(w) + \widehat{\mathcal{R}}(w) - \widehat{\mathcal{R}}(w')\right) && (42)\\
&\leq (1 - \lambda/\beta)\|w - \bar{w}\|^2 && (43)
\end{aligned}
$$

(1): strong convexity, (2): Descent lemma

$\square$

Smooth and strongly convex case

Now let us consider gradient flow:

Theorem 10.
If $\widehat{\mathcal{R}}$ *is* $\lambda - \mathrm{sc}$, *a minimum* $\bar{w}$ *exists, and the GF* $w(t)$ *satisfies*

$$\|w(t) - \bar{w}\|^2 \leq \|w(0) - \bar{w}\|^2 \exp(-2\lambda t) \tag{44}$$

$$\widehat{\mathcal{R}}(w(t)) - \widehat{\mathcal{R}}(\bar{w}) \leq (\widehat{\mathcal{R}}(w(0)) - \widehat{\mathcal{R}}(\bar{w})) \exp(-2t\lambda) \tag{45}$$

*: As in all other rates proved for GF and GD, $\frac{t}{\beta}$ is replaced by t.

Smooth and strongly convex case

To prove Theoerm 10, we need to prove Grönwall's inequality first:

Lemma 11.

Let β and u be real-valued continuous functions in interval $I = [a, \infty)$ or $[a, b]$ or $[a, b)$, if u is differential in the interior I° of I and satisfies the differential inequality:

$$u'(t) \leq \beta(t)u(t), \quad t \in I^\circ$$

then for all $t \in I$ we have:

$$u(t) \leq u(a) \exp\left(\int_a^t \beta(s)\mathrm{d}s\right) \tag{46}$$

Smooth and strongly convex case

Proof of Grönwall's inequality:

Define the function

$$v(t) = \exp\left(\int_a^t \beta(s)\mathrm{d}s\right), \quad t \in I.$$

Note that v satisfies

$$v'(t) = \beta(t)v(t), \quad t \in I^\circ,$$

with $v(a) = 1$ and $v(t) > 0$ for all $t \in I$. By the quotient rule

$$\frac{d}{dt}\frac{u(t)}{v(t)} = \frac{u'(t)v(t) - v'(t)u(t)}{v^2(t)} = \frac{u'(t)v(t) - \beta(t)v(t)u(t)}{v^2(t)} \leq 0, \quad t \in I^\circ$$

Thus the derivative of the function $u(t)/v(t)$ is non-positive and the function is bounded above by its value at the initial point a of the interval I : $\frac{u(t)}{v(t)} \leq \frac{u(a)}{v(a)} = u(a), \quad t \in I$
which is Grönwall's inequality. $\square$

Smooth and strongly convex case

Proof of Theorem 10:

By first-order optimality in the form $\nabla\widehat{\mathcal{R}}(\bar{w}) = 0$, we have:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\|w(t)-\bar{w}\|^2 &= \langle w(t)-\bar{w}, \dot{w}(t)\rangle &(47)\\
&= -\langle w(t)-\bar{w}, \nabla\widehat{\mathcal{R}}(w(t))\rangle &(48)\\
&= -\langle w(t)-\bar{w}, \nabla\widehat{\mathcal{R}}(w(t)) - \nabla\widehat{\mathcal{R}}(\bar{w})\rangle &(49)\\
&\overset{(1)}{\leq} -\lambda\|w(t)-\bar{w}\|^2 &(50)
\end{aligned}$$

where (1) uses an property: f is λ -strongly convex iff
$(\nabla f(x) - \nabla f(y))^T(x-y) \geq \lambda\|x-y\|^2$ $\square$

Smooth and strongly convex case

Proof continued:

By Grönwall's inequality, this implies

$$\begin{aligned} \|w(t)-\bar{w}\|^2 &\leq \|w(0)-\bar{w}\|^2 \exp\left(-\int_0^t 2\lambda \mathrm{d}s\right) &(51)\\ &\leq \|w(0)-\bar{w}\|^2 \exp(-2\lambda t) &(52)\end{aligned}$$

which prove the first part. As for the objective function part:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}(\widehat{\mathcal{R}}(w(t))-\widehat{\mathcal{R}}(\bar{w})) &= \langle \nabla\widehat{\mathcal{R}}(w(t)), \dot{w}(t)\rangle &(53)\\ &= -\|\nabla\widehat{\mathcal{R}}(w(t))\|^2 &(54)\\ &\overset{\text{(PL)}}{\leq} -2\lambda(\widehat{\mathcal{R}}(w(t))-\widehat{\mathcal{R}}(\bar{w})) &(55)\end{aligned}$$

Smooth and strongly convex case

Proof continued:

By Grönwall's inequality, this implies

$$\widehat{\mathcal{R}}(w(t)) - \widehat{\mathcal{R}}(\bar{w}) \leq (\widehat{\mathcal{R}}(w(0)) - \widehat{\mathcal{R}}(\bar{w})) \exp(-2t\lambda) \tag{57}$$

we finish the proof now. $\square$

- Thanks for your time!
- Next time we will discuss stochastic gradients.